Aids Policy and Psychology: A Theoretical Approach*

Andrew Caplin and Kfir Eliaz†
Economics Department and Center for Experimental Social Science
New York University

May 2002
Preliminary and Incomplete

1 Introduction

Economic theorists have proposed many schemes to reduce the damage caused by externalities. How best to slow the spread of AIDs would seem to be an important case in point. Yet despite the pioneering efforts of Philipson and Posner [1995] (henceforth P&P) and Kremer [1996], economic theorists have largely ignored this question. Indeed they have given scant attention to *any* health-related externalities, despite their profound social importance.

Why have medical questions received so little attention from economic theorists? One possible reason is that health-related decisions are psychologically complex. The critical role of psychology is hinted at by P&P when they discuss the role of *fear* in limiting the efficacy of certain AIDS policies. In particular, they discuss the potential impact of verifiable "HIV-cards" that offer proof to all that one does not have the virus. In an idealized version of this scheme, they speculate that there would be assortative matching, with those who were were certified not to be infected matching only with others of their type. Yet, following the empirical findings of Lyter et. al [1987], they argue that not many would be willing to take such a test for psychological reasons:[1]

> "many people are fearful of tests which may show they are doomed even if the probability of that result is very low" [P&P, p. 472].

In this paper we present a theoretical approach to AIDS policy that explicitly incorporates just this form of fear. We use our approach to reassess the potential for certification policies to

*We thank Alessandro Lizzeri, Ronny Razin and Andrew Schotter for their valuable comments. We also thank the C.V. Starr Center of NYU for financial support.

†269 Mercer St., New York, NY 10003. E-mails: andrew.caplin@nyu.edu, kfir.eliaz@nyu.edu

[1]Lyter et. al find that many patients are reluctant to learn their HIV status even when confidentiality is guaranteed. Among the most common reasons cited for rejecting this information is precisely the anticipation of severe psychological distress if the result is positive.

reduce the spread of the disease. We outline circumstances in which a variant of the HIV-card scheme may be a very effective policy tool, even when fear is profound.

We begin in section 2 of the paper by developing a strategic model of the spread of AIDS, and the potential role of certification in limiting its spread. We confirm the conjecture of P&P: in the absence of fear, there is a unique equilibrium in which all agents test, and matching is assortative. In section 3 we incorporate into the model a fear-induced preference for late resolution of uncertainty, and confirm that it may indeed render the certification policy ineffective. If fear is sufficiently important, there is an equilibrium in which no-one tests, and the disease continues to spread.

What can be done to reduce the spread of AIDS? We consider a class of message transmission policies more general than the HIV-card scheme of P&P. The policy maker is able both to assess the health status of private agents, and to pass back certifiable messages to them based on the results of this assessment. These messages can have a more subtle interpretations than the "either you have it or you don't" message implied by the simple HIV-card scheme.[2] We explore the impact of these general message transmission policies on equilibrium outcomes. As in implementation theory, we look for policies such that the *unique* equilibrium outcome in the subsequent matching game is as beneficial as possible. Our focus on uniqueness reflects uncertainty as to which equilibrium will arise in the face of multiplicity. If an undesirable equilibrium (e.g. not testing) is the status quo, then the presence of some other more attractive equilibrium is irrelevant, since it will never be reached.

Sections 5 through 7 explore the details of policy design in an important special case. We assume that the policy maker's primary goal is to reduce the rate of infection. We then search for conditions under a policy exists guaranteeing that there will be *no* new infections. Section 5 proves that we can restrict attention in this search to simple policies, somewhat analogous to the certification scheme of P&P. Sections 6 and 7 provide more detailed analyses of such policies under two different assumptions concerning feasibility. Section 6 explores relatively simple "unconditional" policies that are unable to respond to the fine structure of the test results. Section 7 considers conditional policies that are fully flexible. In both cases we provide reasonable conditions under which the policy maker is able to entirely prevent new infections, even in the presence of fear.

If policies can be programmed into a machine, sections 5 through 7 contain a complete resolution to the externality problem in our model. Yet the situation is different if the policy maker is instead a physician, as in Caplin and Leahy [1999]. In fact in section 8 we show that the policies outlined in earlier sections are then not credible. Those policies are explicitly designed to limit the negative information provided to those who are actually infected, in order to induce fearful individuals to test. Unfortunately this requires pooling a certain number of healthy and unhealthy individuals (i.e. providing them with the same message). We show that it is not credible that a caring physician with a healthy patient would agree to pass on the ambiguous message called for by our policies.

[2] Given our fundamental interest in psychological incentives, we do not consider variations in monetary incentives of the type analyzed by P&P.

It is important to end with a disclaimer. Our model of AIDS transmission is highly stylized. Our solution to the externality problem is specific to the model, and is very far from being ready for use in the field. Yet we have great faith in the underlying message. Rather than viewing psychological realities as a barrier to economic analysis, we believe that they are profoundly enriching. The time has come not only to acknowledge their importance, but also to incorporate them into policy analysis.

2 Assortative Matching

We first outline a game of testing and matching, Γ, that defines the social context underlying the spread of the disease. Society consists of a fixed finite set of individuals. We focus in particular on the three player case, since this is the smallest number in which competitive forces can come into play to induce testing. At the start of the game, each individual can be in one of two possible health states, infected or not infected, and there is a common initial probability p of being infected. There is a testing technology available, which is assumed to be costless and definitive. A positive test result implies that the subject is surely infected, a negative result that he is surely not. We assume also that there is a certification technology associated with the test: a player who tests negative receives an official (non-falsifiable) document that confirms his test result.

The game itself is played in two stages. In the first stage each individual has the options T and NT, corresponding to the decisions to test and to not test respectively. All players make their testing decisions simultaneously, and these decisions are not observed by other players. The second stage of the game involves all three individuals meeting up and forming at most one mutually acceptable pairing. We model the matching stage as follows:

- When the three players meet, those who tested negative simultaneously decide whether or not to publically reveal their certificates. For an individual who does not show a certificate, inferences about whether or not he took the test, and the result of the test if taken, are based on strategies alone.

- At the start of the matching stage, nature picks one of the three players to be first proposer, $P \in N = \{1, 2, 3\}$. This player gets to make a take-it-or-leave-it matching offer based only on the messages revealed by the remaining two players. Having seen these messages, player P can publically announce one of the four statements: C, $\emptyset$, $\{C \cup \emptyset\}$ or Q, standing respectively for a willingness to match with a certified player, a player passing no message, either player, and a desire to quit the game. If he chooses to make an offer, there must be at least one player in the game who satisfies the stated criterion.

- At this stage, nature picks one of the other players to be first responder, FR, leaving the remaining player as second responder, SR. Precisely how this selection is carried out depends on what the proposer stated. If P quit, then the remaining two players are equally likely to be first responder. If P made any of the other three statements, then FR is randomly picked from among those who satisfy the stated condition.

- If P quit, FR gets a yes/no choice between either making an offer to SR or quitting, in which case the game ends without a match. If P proposed a match, FR has the choice

either to accept or reject the offer. If he accepts, the game ends with the agreed upon match taking place. If he rejects, he again has the choice between either making an offer to SR and quitting to end the game.

- If the game reaches him, SR simply gets to say yes or no to the offer from FR. An affirmative answer results in the match taking place, while a negative results in the game ending without a match.

At the end of the round of proposals, either there is some agreeable match or there is not. The no match utility is normalized to zero. The match itself yields direct utility normalized to one to each partner. On the other hand, if a healthy individual is matched with an unhealthy individual, he ends up unhealthy himself. The disutility of bad health is $H > 0$. Decisions are made to maximize the expected benefits of matching, net of any additional health costs incurred.

The above game is designed to satisfy somewhat complex informational requirements. On the one hand, we want players potentially to be perfectly informed about the quality of the pool of available partners, to ensure that there is an advantage to testing. It is for this reason that players have the option of revealing their certification status. Yet unless we are careful, this form of revelation opens up strategic possibilities that seem artificial given the somewhat anonymous settings we have in mind. For example, if player 2 is believed to be testing but player 3 is not, and if neither of them shows a certificate, then player 3 is a preferable partner in the eyes of player 1. His failure to show a certificate is uninformative about his disease status, while player 2's may be interpreted as revealing infection. While such inferences are possible in our model, they are not so easy in the somewhat anonymous settings we have in mind. To restore anonymity, we therefore restrict players to strategies in which they can select partners *only* on the basis of their revealed certification status, rather than their identities.[3] While player 1 prefers to match with player 3 in the above example, his offer to match with an uncertified player is equally likely to be delivered to player 2. He must pick based on the his view of the larger population of uncertified individuals.[4]

Our game has an extensive form with imperfect information but perfect recall (see Osborne and Rubinstein [1994], pp. 200-201). The appropriate solution concept is *sequential equilibrium*[5](Kreps and Wilson [1982]). This concept consists of both a strategy profile β and a belief system μ (which together constitute an *assessment*), where a belief system specifies, for each

[3]Another reason for restricting strategies in this way is to remove asymmetric equilibria in which one of the players gets singled out for punishment should he be of equivalent status to the other.

[4]An equivalent way to model anonymity in the matching game is the following. Before the matching game begins each player who has a certificate puts his certificate in a box. The players then get to observe the box, so that each only knows the number of certificates that were revealed, but does not know the identity of their owners. Before each player makes a proposal he must take his certificate out of the box and make his offer. Thus, when making his offer, P/FR does not know the identity of any other player who may have put a certificate in the box. However, everyone else knows that P/FR was certified and that one of the certificates in the box is his. This alternative way of modeling the revelation game leads to exactly the same set of matching strategies that are possible in the model that we proposed. However, our modeling approach is simpler: we eliminate the act of taking one's certificate from the box when making an offer.

[5]Note that we cannot use the notion of perfect Bayesian equilibrium, since the testing decisions are not observable (see Osborne and Rubinstein [1994], p.231).

information set, the beliefs held by the players who have to move at that information set about the history that occured. In particular, a player's equilibrium beliefs must assign a probability distribution over the set of histories that led to an information set that should not have been reached had the players followed their equilibrium strategies.

We are now in a position to verify the conjecture of P&P that certification can result in an equilibrium in which no new infections occur. In doing this, we impose bounds on H to ensure that both a healthy individual and an untested individual will match with an untested individual, but that an untested individual will not match with an unhealthy one. Before provind the result we introduce some standard notational conventions. An information set for player i will be denoted by I_i. We denote the set of actions available to player i at the information set I_i by $Y(I_i)$. The probability that the behavioral strategy β_i at the information set I_i assigns to the action $y \in Y(I_i)$ is denoted $\beta_i(y \mid I_i)$.

Proposition 1 *Assume that $p < \frac{1}{2}$, and that $H \in [\frac{1}{1-p}, \frac{1}{p}]$. Then the unique sequential equilibrium outcome involves assortative matching. All players test, and matches form among players who are already in the same state: healthy with healthy, unhealthy with unhealthy.*

Proof. We proceed in two steps. In the first step we prove existence: we describe an assessment that leads to assortative matching, and show that this assessment is an equilibrium. In the second step we prove uniqueness.

Step 1: Existence. Consider an assessment (β^*, μ^*) that satisfies the following.

The behavioral strategy profile β^*: Let I_i^0 be the initial information set of player i (following the null history). Then for each player i: $\beta_i^*(I_i^0)$ assigns probability one to T. Any player who tested and received a certificate reveals his certificate.

The remaining information sets of each player can be partitioned into three sets: (1) the set $\mathcal{I}_i^{TNC}$, which consists of all information sets at which player i arrives if and only if he tests but is not certified, (2) the set $\mathcal{I}_i^{TC}$, which consists of all information sets at which player i arrives if and only if he tests and is certified and (3) the set $\mathcal{I}_i^{NT}$, which consists of all information sets at which player i arrives if and only if he does not test.

For each i and for each information set $I_i \in \mathcal{I}_i^{TNC}$ the behavior dictated by $\beta_i^*(I_i)$ depends only on i's role in the matching subgame (i.e. whether i has been selected as P, FR or SR):

- If selected as P, choose $\{C \cup \emptyset\}$.
- If selected as FR, accept any offer that P has made. If P quit, make an offer to SR regardless of certification status.
- If selected as SR, accept any offer.

For each i and for each information set $I_i \in \mathcal{I}_i^{TC} \cup \mathcal{I}_i^{NT}$ the behavior dictated by $\beta_i^*(I_i)$ depends not only on i's role in the matching subgame, but also on the certification status of the other players:

- If selected as P, select C if there is one, otherwise quit.
- If selected as FR, accept an offer from a certified P, reject otherwise. If P quit or the offer from FR is rejected, make an offer to SR only if SR is certified.
- If selected as SR, accept an offer only if FR is certified.

The belief system μ^*: At any information set in the matching subgame each player assigns probability one to the following history: all players have tested, any player who is uncertified has AIDS with certainty while any certified player is healthy with certainty.

It is immediate that the behavior dictated by β^* in the matching subgame produces the assortative matching outcome. Whatever test outcome is realized by a majority of the population determines the nature of the match. If there are two or more who are not infected, then it is certain that a pair of them will match. If there are two or more who are infected, it is certain that a pair of them will match. Overall, matching is guaranteed to take place at no cost in terms of increased infection. It remains to show that (β^*, μ^*) is an equilibrium of Γ.

To establish that (β^*, μ^*) is an equilibrium we need to show that this assessment is a sequential equilibrium of Γ. We begin by verifying that (β^*, μ^*) is sequentially rational.

Sequential rationality

Consider the information sets of the players in the matching subgame. Given our payoff assumptions, an infected individual always benefits from matching. However, players who are either healthy or untested will note that since lack of certification implies infection in the equilibrium, uncertified matches are too much of a health risk. With this, FR if he is unhealthy knows that if he rejects an offer from P, he will match only if SR is also unhealthy, hence he will be willing to accept any offer. On the other hand, a healthy or untested FR is absolutely unwilling to match with an uncertified P, and has nothing to gain by rejecting an offer from a certified proposer. It is then clear that both a healthy and an untested P will make proposals only to a certified player, while an unhealthy one will propose to everyone.

Given the players' matching strategies, it is weakly dominating to reveal one's certificate. It remains to show that choosing T with probability one at the start of the game is sequentially rational. Given player i's beliefs μ_i^* and the behavioral strategies of the other two players β_{-i}^*, by testing player i obtains positive expected utility from matching, while suffering no increase in the probability of infection: if he is found to be uninfected, he will not match with anyone unless they too are uninfected. In contrast, if player i deviates by not taking the test, he is never able to match: he gets an offer only if the other two players are unhealthy, and in turn he rejects any such offer. It is therefore strictly advantageous to test.

Consistency

To verify consistency of (β^*, μ^*) we perturb it as follows. For each player i and for each information set I_i we assign probability ε to any feasible action not chosen according toβ^*, while assigning the complementary probability to the pure strategy $\beta^*(I_i)$. Now beliefs at all information sets can be computed using Bayes' rule. It is immediate that the probability assigned to the event that a certified player is healthy conditional on taking the test is one. In addition, the

probability assigned to the event that a noncertified player is infected conditional on him taking the test (respectively, not taking the test) remains one (respectively, p). Clearly, the perturbed strategy profile and the associated belief system converge to (β^*, μ^*) as $\varepsilon \to 0$, confirming consistency.

Step 2: Uniqueness. Assume there exists an equilibrium assessment (β, μ) that does not generate assortative matching. Suppose β is such that each player tests for sure. Then for assortative matching to fail, a certified player must be matching with a noncertified one. This means that β satisfies that a certified player agrees to match with a noncertified player. Moreover, by the consistency of μ, the beliefs of each player must satisfy: (1) All players have tested, and (2) a player who tested and was not certified is infected for sure. This implies that β is not sequentially rational since given μ, a certified player will not want to match with a noncertified player.

From the argument above, it follows that (β, μ) must satisfy that at least one player (say, player 1) does not test with positive probability. If β requires player 1 to randomize between testing and not, then given his belief system and the assessments of the other players, player 1 must be indifferent between T and NT. This implies that in order to show that β is not sequentially rational, it suffices to show that testing with certainty is a profitable deviation for player 1.

Suppose player 1 deviates from β by testing with certainty, and yet fails to be certified. At any history in which player 1 is not certified, players 2 and 3 continue to hold their initial beliefs (that is, they do not update their belief and assign probability one to the event that player 1 is infected). This means that when player 1 tests with certainty but fails, his expected payoff from matching in the matching subgame is at least equal to his expected payoff before the deviation (since player 1 can "hide" his noncertification). The only difference is that he may now accept an offer that previously he rejected, a strict net gain. This means that it suffices to show that when player 1 is not infected, tests, and is certified, his expected payoff is strictly higher than when he does not test.

Consider then the case in which player 1 deviates to testing, is certified as healthy, and chooses to reveal his certificate. We show that this is always a strictly profitable deviation in comparison with being healthy and not testing. A key observation is that the consistency of μ implies that players 2 and 3 must believe in this case that player 1 surely healthy, and therefore always include him in any matching offer. The end result is that player 1 always benefits from this, although precisely how this benefit comes about depends on what other players do and on the state of the world.

- If either of the other players reveals a certificate, the benefit to the deviation to testing and revelation is a strict increase in matching probability with no additional infection risk.

- If neither of the other players reveals a certificate, and player 1 is selected as proposer, he cannot lose from the deviation: if he is interested in making an offer given the probable infection status of the other two players, he so offers, otherwise he withdraws. This is simply a better informed decision than before, and cannot involve any loss. If he is not selected as proposer, then he is sure to be included in any offer that he is interested in

accepting. If he is not, or if he chooses to turn the offer down, he is doing so because this match is strictly unworthwhile for him. If this is the case, then it is certain that knowing that he was healthy, he would have wanted to turn down an offer at an equivalent stage in the game in which he did not reveal his certificate, since failure to make such an offer can only convey good news about the type of the individual making the offer.

- To confirm that deviation to testing and revealing in the good state is strictly better than not testing, note from the above that it is strictly better when one of the players reveals, weakly when neither does. This leaves only the case in which the probability that neither of the other players reveals a certificate is a probability 1 event. Strict gains occur in this case because of the desire to match with a random untested member of the population (A1).

We have shown that any assessment that does not lead to assortative matching has the property that the behavioral strategy of at least one of the players is not sequentially rational. In particular, there exists no equilibrium in which some player does not test with certainty. This completes the proof of Step 2 and also completes the proof of Proposition 1.■

Proposition 1 confirms the conjecture of P&P in a setting in which there is no fear of the outcome. We now turn to the second part of their conjecture, which asserts that this beneficial outcome is not assured if "many people are fearful of tests which may show they are doomed even if the probability of that result is very low".

3 Anxiety as a Barrier to Testing

To capture reluctance to learn, we add an increasing and convex "anxiety" function, $A(\pi)$, where the π argument reflects beliefs after the matching stage is over. We assume separability of the anxiety argument from the matching and health-related arguments in the utility function. From a decision-theoretic perspective, this reduces to assuming a preference for late resolution of uncertainty in the sense of Kreps and Porteus [1978]. However the credibility analysis of section 6 require us to take a stand on the separate impact of beliefs and outcomes in influencing utility, as in Caplin and Leahy [1999]. It is for this reason that we interpret the A function as specifying the realized level of anxiety as it depends on prior beliefs, regardless of whether or not these beliefs are accurate in light of any test that has been taken.

Rather than work with a completely general A function, we pick the following functional form designed for flexibility and tractability:

$$A(\pi) = \begin{cases} \frac{K}{(1-p)^r}\,(\pi - p)^r & for \quad \pi \in [p,1] \\ 0 & for \quad \pi < p \end{cases}$$

The parameter $r > 1$ captures the degree of convexity of this function, while the parameter $K > 0$ captures the maximum level of anxiety, which is associated with the certainty that one will end up infected. Setting anxiety to zero below p simplifies formulae without changing essentials.

What conditions on K and r ensure existence of an uninformative equilibrium? As in Proposition 1, we need a healthy individual to be willing to match with an untested one. However the

additional anxiety term means that this condition by itself no longer guarantees that two untested individuals will agree to match. Ensuring that this holds is one part of Axiom 1 (henceforth A1), which ensures existence of a no-testing equilibrium.

Axiom 1 *We assume that a healthy individual will match with an untested one,* $1 - pH > 0$. *We assume also that the anxiety parameter* K *satisfies the inequality,*

$$\frac{(1-p)(1-pH)}{p} \leq K \leq \frac{1-pH(1-p)}{p^r};$$

Note that the range of possible K values in A1 is always non-empty. In particular, $K = \frac{(1-pH)}{p}$ is a specific value that satisfies both conditions in A1.[6] Note also that the set of values of K satisfying the inequality increases with r: the more convex is the anxiety function, the easier it is for great fear of learning the truth to be combined with little discomfort at remaining ignorant and continuing to take small risks. In fact the upper bound heads to infinity as r increases.

To precisely interpret the upper bound on K in A1, it is helpful to rewrite it as follows,

$$p^r K \leq 1 - pH(1-p).$$

The left hand side of this inequality is the incremental anxiety caused when an untested player matches with another untested. The right hand side is the gain from trade in such a match, taking into account the incremental health risk. This upper bound on K ensures that the anxiety loss in such a match is dominated by the gains from trade. If this condition failed, there would be no new infections even without testing. The precise interpretation of the lower bound on K in A1 is somewhat more intricate, although its role is straightforward: it ensures that testing and failing is such an unpleasant prospect that a universal rejection of testing is a stable state. Note that if the test is taken, the ex ante expected pure anxiety cost of a negative test result (infected) is pK. A1 places a lower bound on just this quantity,

$$pK \geq (1-p)(1-pH).$$

The right hand side of this inequality is the ex ante expected net gain if the test is taken, the result is negative, and there is then a certain match with an untested player. If expected anxiety costs exceed these expected gains, then non-testing is indeed an equilibrium, as confirmed in the following proposition.

Proposition 2 *With axiom 1, there exists an equilibrium in which no one tests, and in which there is always a match achieved.*

P roof. The first role of the axioms is to structure the outcome conditional on no players testing. The upper bound on K in A1 precisely implies that in this setting, players will willingly match. Such a match impacts utility by adding the matching benefit, but subtracting the

[6] The case $K = \frac{(1-pH)}{p}$ has a straightforward interpretation: it implies that the anxiety loss from a bad test result, pK, is equal to the gain from trade if certified clean, $1 - pH$. More broadly, note that the lower bound in A1 forces up the expected anxiety loss close to this gain from trade.

increased health cost and the increased anxiety. Substitution in the anxiety function yields the condition for this overall net benefit of matching to be positive as,

$$1 - Hp(1-p) - \frac{K}{(1-p)^r}\,[p(1-p)]^r \geq 0.$$

Rearranging this expression reduces it to the upper bound on K in A1.

The second role of A1 is to bound the benefits of deviating to testing. When we ignore anxiety, we know that the gain in expected utility from a deviation to testing is $\frac{1}{3}(1-p)(1-pH)$: when the individual does not match in the 3 player game (probability $\frac{1}{3}$) and he tests clear, his probability of being matched now rises to 100%. With this incremental probability of $\frac{1}{3}(1-p)$ he gains the match utility of unity, but loses the health cost of switching to bad with probability p. The cost of this gain is the increase in anxiety that comes from learning one's status. Without testing, the belief argument of the anxiety function is a $\frac{2}{3}$-$\frac{1}{3}$ mix of $p + p(1-p)$ and p. With testing, the only time the argument of the anxiety function exceeds p and therefore produces strictly positive anxiety is when the agent is in fact found to be infected. Noting the functional form assumption on the anxiety function, the overall condition in which no one breaks the fully ignorant equilibrium is that the anxiety cost be at least as big as direct utility gain,

$$pK - \frac{2Kp^r}{3} \geq \frac{1}{3}(1-p)(1-pH)$$

and

$$K \geq \frac{(1-p)(1-pH)}{3p - 2p^r} \qquad (*)$$

With A1, we know that K does indeed satisfy this inequality, since it is bounded below by $\frac{(1-p)(1-pH)}{p}$ which exceeds the right hand side in $(*)$. ■

4 Policy Analysis

4.1 The General Framework

In this section we extend the power of the policy maker by allowing her to pass on more sophisticated messages than merely the result of the AIDS test. We begin by formally laying out the full set of message-transmission options available to her. We refer to each such message transmission policy as a mechanism.

To define the general class of mechanisms, let t denote a profile of test results. t is assumed to be a random variable whose range is $T = \{+, -, NT\}^3$, where $(+)$ stand for a postive test result (i.e., AIDS carrier), $(-)$ stands for negative test result (i.e., not an AIDS carrier) and NT stands for "not testing". The random variable t is a function of the individuals' testing decisions at the start of the game. For an individual who tested, t is either $(+)$ or $(-)$ with probabilities p and $1-p$. For an individual who did not test, $t = NT$ with certainty.

Let M be some set of messages. A mechanism is a function that assigns to each possible test result a probability distribution over the set $\mathcal{M}$,

$$\mathcal{M} = (M \cup \{\emptyset\}) \times (M \cup \{\emptyset\}) \times (M \cup \{\emptyset\})$$

where $\emptyset$ denotes "no message" (also called the null message). We assume that for all test results, player i directly receives only his own message: he is left to observe or to infer the messages received by others in the play of the game.

We assume that the mechanism the planner commits to is common knowledge. Thus, a mechanism G induces an extensive game $\Gamma(G)$ with imperfect information and perfect recall. The nature of this game is little impacted by the replacement of the AIDS certificate with the more general set of messages outlined above. As before, the first decision each player makes is whether or not to show up and be "tested" by the mechanism. Upon receiving their messages, all players meet and decide whether or not to reveal these messages to one another. Nature then selects the first proposer, P, who either withdraws, or makes an offer to some non-empty subset S of M. FR is chosen as before, and decides whether or not to accept any offer that P might have made. If he rejects or if he did not receive an offer from P, he has the option of making his own offer to SR.

While our "message-transmission" approach is borrowed from implementation theory, our policy problem is quite different. One difference is that in implementation theory, the policy maker is at an informational disadvantage. Private agents know their own type, the policy maker does not. In our model the informational imbalance is reversed: only the planner has the technology to identify private agents' types. A second difference is that in implementation theory, the policy maker is free to design a game in which the players must participate, and whose rules they must obey. In contrast, in the AIDS setting, the planner cannot force players to "join in" the certification stage of the game, nor can she directly influence their behavior in the larger social world once they have tested. Her only power lies in her ability to assess private states, and to pass back messages to the private agents based on these assessments. For such policies to have any effect, some of the players must find it worthwhile to test, and some of those must in turn choose to display their messages in the matching stage of the game.

4.2 Conditional and Unconditional Mechanisms

We will be interested in two different classes of mechanism: those that respond differently in each state of the world, and those that have less flexibility to respond contingency by contingency. We refer to the broader class as the set of conditional mechanisms, and the narrower class as the unconditional mechanisms. To formally define these classes, let $g_i(m \mid t)$ denote the probability that player i receives the message $m \in M \cup \{\emptyset\}$ conditional on the test result t. Our first global assumption asserts that no message can be given to an individual who does not show up for the test.

$$(G1) \textit{ untested individuals receive no message: } t_i = NT \Longrightarrow g_i(\emptyset \mid t) = 1$$

Assumption $G1$ ensures that a player with an unattractive message can pretend not to have tested (in equilibrium players may infer that this is likely to be a pretense, and update beliefs accordingly). For unconditional mechanisms, this is all that is assumed

Definition 1 *The set of conditional mechanisms are precisely those satisfying* $(G1)$*. We denote the set of conditional mechanisms by* $\mathcal{G}^C$*.*

A conditional mechanism can respond differently to an infected individual depending on the infection status of surrounding players. In practice, it may be infeasible to display this form of flexibility. We therefore analyze also mechanisms that satisfy additional equal treatment and independence requirements.

($G2$) *Equal treatment of test results*: $t_i = t_j \Longrightarrow g_i(m \mid t) = g_j(m \mid t)$ for all $m \in M \cup \{\emptyset\}$.

($G3$) *Independence*: $t_i = t_i' \Longrightarrow g_i(m \mid t) = g_i(m \mid t')$ for all $m \in M \cup \{\emptyset\}$ and $t, t' \in T$.

Definition 2 *The set of unconditional mechanisms are precisely those satisfying* $(G1) - (G3)$. *We denote the set of unconditional mechanisms by* $\mathcal{G}^U \subset \mathcal{G}^C$.

There are several cases in which unconditional mechanisms may be more reasonable than unconditional mechanism. Equal treatment, $G2$, is needed if there are exogenous constraints (legal, moral or political) that prevent the planner from discriminating between two individuals with identical test results. Independence is valuable in cases in which tests are taken in sequence, and one does not wish to wait for all tests to be taken before handing out the results of earlier tests.

4.3 The Policy Maker's Objective

What makes a particular mechanism attractive? The policy maker is concerned with three outcomes.

1. The probability that a new player is infected, $\pi_I \in [0, 1]$.
2. The probability that two players match, $\pi_M \in [0, 1]$
3. The average anxiety level of the players, $\bar{A} \in [0, K]$.

To capture the importance of getting rid of infections, we make the specific assumption that the planner has lexicographic preferences. Minimizing infections is her first goal, maximizing gains from trade her second, and reducing anxiety the third.

As in implementation theory, we restrict attention to mechanisms for which the ultimate outcome is deemed to be "predictable". We use the same definition of predictability as in implementation theory: the game $\Gamma(G)$ must have a unique pure strategy equilibrium outcome $(\pi_I(G), \pi_M(G), \bar{A}(G))$ in the payoff space[7]. Within this class of mechanisms, the policy maker selects according to her lexicographic social welfare function.

[7] This conception follows the tradition of standard models in implementation theory, in which mixed strategy equilibria are ignored (Jackson [2001], pp. 682-685).

5 A Special Case: No Infections

5.1 A Parameter Restriction

Under what circumstances can the policy maker design a mechanism that ensures no infections, even when A1 is valid and fear acts as a disincentive to testing?

Definition 3 *A mechanism $G \in \mathcal{G}^U$ (resp. $\mathcal{G}^C$) implements zero infections if $\Gamma(G)$ has a unique pure strategy equilibrium outcome $(\pi_I(G), \pi_M(G), A(G))$, with $\pi_I(G) = 0$.*

Existence of such a mechanism requires assumptions on the parameters of the model in addition to A1. As we will show in the sections that follow, A1 in combination with Axiom 2 (henceforth A2) specifies the entire bank of conditions that between them ensure that one can implement zero infections using either a conditional or an unconditional mechanism. As will be seen, these conditions are far from necessary.

Axiom 2 *We assume that $H > 4$, that $p \leq \frac{H-1}{H^2}$, and that,*

$$\left(\frac{2}{H}\right)^{r-1} \leq \frac{(1-pH)^2}{3pK}.$$

The assumption on H is transparent and probably unobjectionable: AIDS would not be much of a disease if the costs were not at least four times as high as the benefits of a match. The assumption on p is also of limited impact: A1 already asserts that $p \leq \frac{1}{H}$, so that this slightly tighter bound does not seem particularly harsh. The final inequality is essentially a convexity assumption. The left hand side of the inequality diminishes to zero as r increases, while the right hand side is a number that is below 1 (in light of A1), but is independent of r.

What are the implications of combining A1 and A2? We have seen that A1 is satisfied if $pK = 1 - pH$.Note that in this case A2 simplifies to,

$$\left(\frac{2}{H}\right)^{r-1} \leq \frac{(1-pH)}{3}$$

For example, note that if the health costs are significant relative to the matching benefits, say $H = 12$, and if health risks by themselves reduce by 50% the gains from trade, $1 - pH = \frac{1}{2}$, this inequality is satisfied for $r \geq 2$.

5.2 Zero Infections and Simple Mechanisms

Before proving specific results concerning the structure of optimal conditional and unconditional mechanisms, we look first to simplify the set of mechanisms to be analyzed. Our general class is far broader than the certification policies discussed by P&P. Yet a great deal of structure is added when we focus on the goal of implementing zero infections. Proposition 3 shows that the best way to do this requires at most two distinct messages (in addition to the null message).

Definition 4 *The set of **simple** mechanisms are all those in which the message space is a doubleton.*

Definition 5 *The set of* ***very simple*** *mechanisms are all those in which the message space is a singleton.*

We let $\mathcal{G}^{SU} \subset \mathcal{G}^{U}$ $\left(\mathcal{G}^{VSU} \subset \mathcal{G}^{U}\right)$ denote the class of (very) simple unconditional mechanisms, and $\mathcal{G}^{SC} \subset \mathcal{G}^{C}$ $\left(\mathcal{G}^{VSC} \subset \mathcal{G}^{C}\right)$ denote the class of (very) simple conditional mechanisms. The following result establishes that in the case of zero-infection mechanisms, any outcome that can be achieved with more intricate message spaces can equally be achieved with a simple mechanism.

Proposition 3 *With A1, if there is any mechanism $G \in \mathcal{G}^{U}$ (resp. $\mathcal{G}^{C}$) implementing zero infections, then there is a simple mechanism $G' \in \mathcal{G}^{SU}$ (resp. $\mathcal{G}^{SC}$) that also implements zero infections, and does at least as well in terms of the planner's objective. In the unconditional case, the mechanism can be very simple.*

Proof. We provide a direct proof only for the unconditional mechanism. The proof for the conditional mechanism is constructive, and is presented in Proposition 5 below.

For there to be an unconditional mechanism $G \in \mathcal{G}^{U}$ implementing zero infections, it must have four properties:

1. All players who test and receive "good" messages (at least as likely conditional on not being infected) must in fact be free of disease.
2. Those who receive good messages must be unwilling to match with those who do not.
3. Those who do not receive good messages must be unwilling to match with one another, unless they themselves receive a message perfectly predictive of their being infected.
4. Failure to take the test must result in certain rejection in the matching game.

To confirm this, note first from A1 that all will be willing to match with those who get and display positive messages: in particular they will be willing to match with one another. To ensure that such matches do not create disease, they must be between two individuals who are disease free, confirming (1). Now this implies also (2), since there is surely risk of infection when a disease free individual matches with an individual who did not receive good news, and therefore clearly has a positive probability of being infected. (3) follows, since unless non-reception of a positive message implies that one has the disease for sure, matches among two without a positive message also run the risk of creating infection. (4) is globally obvious for getting zero infections.

What this implies is that there are at most three distinct types of player in the ultimate matching game if all test: those who are certified as free of disease (type 1), those who are guaranteed infected (type 2), and those who are given neither of the above certificates, and who find themselves entirely locked out of trade (type 3). We now verify that the best way for the planner to get these three different types of players is to leave undifferentiated the middle group.

Let $G \in \mathcal{G}^{U}$ be some general mechanism that implements zero infections. Let (β, μ) be a sequential equilibrium of $\Gamma(G)$. Let M_i denote the set of messages received in that equilibrium

by players of type $i \in \{1,2,3\}$. Let x and y denote the proportion of players that are of type 1 and 2 respectively in the equilibrium (β, μ). By the consistency of μ, a player who receives a message in M_1 must be believed to be healthy for sure. Similarly, any player with a message in M_3 must be believed to be infected for sure. For G to implement zero infections the equilibrium (β, μ) must satisfy the following:

1. Any player who receives a message $m \in M_1$ agrees to match only with a player who received some message in M_1 (it need not be the same exact message).

2. Any player who did not receive a message in M_2 does not want to match with any player who did receive a message in this set.

3. Any player with a message in M_3 does not want to match with any player whose message is in M_2 or M_3. This means that a player who receives some message in M_3 will not want to match with any player who is believed to be infected with probability $\frac{p}{1-x}$.

4. All players test.

Consider the following simple unconditional mechanism G^{SU}. For any player who tests positive a lottery is held: With probability $\frac{y}{p}$ that player receives the message UC, while with the complementary probability he receives no message. Similarly, for any player who tests negative a lottery is held: With probability $\frac{x}{p}$ that player receives the message C while with the complementary probability he receives no message.

Claim 1 $\Gamma\left(G^{SU}\right)$ *has a unique sequential equilibrium outcome in which all players test, the only players who match are those who received a message and a match occurs between two players with the same message.*

Proof of Claim 1. We proceed in two steps. In Step 1 we characterize the sequential equilibria of $\Gamma\left(G^{SU}\right)$. Step 2 consists of verifying that those are the only two equilibria.

Step 1: *The sequential equilibria.*
Let β^* be a behavioral strategy profile with the following properties.
(S1) Each player tests.
(S2) Only players who receive the message C reveal their message.
(S3) Any player who receives the message C agrees to match only with a player who has the same message.
(S4) Any player with the message UC agrees to match with any player.
(S5) Players who received no message agree to match only with those who received the message C.

Let μ^* be a belief system that is derived from β^* using Bayes rule. At information sets that cannot be reached if β^* is followed, the belief system satisfies the following.
(B1) Any player with the message C is believed to be healthy for sure.
(B2) Any player with the message UC is believed to be infected for sure.

(B3) Any player with no message who offers a match to players with any message in M is believed to be infected for sure.

(B4) Any player with no message who offers a match only to players with the message C is believed to be infected with probability $\frac{p-y}{1-y-x}$.

We now verify that (β^*, μ^*) is indeed a sequential equilibrium. Clearly, the behavioral strategies of the players with non-null messages are sequentially rational. As for players with the null message, note that in (β, μ) a player who received a message in M_3 will not match with a player who receives a message in $M_3 \cup M_2$ who is therefore infected with probability $\frac{p}{1-x}$. Since (β, μ) is a sequential equilibrium, it follows that the behavioral strategy prescribed by β^* to players with the null message is sequentially rational.

To verify the consistency of μ^* we perturb β^* as follows. For each player and for each of his information sets we assign a probability of ε to any action that is not chosen according to β^*. As $\varepsilon \to 0$, the belief system that is derived from the perturbed strategies using Bayes' rule converges to μ^*. This established that (β^*, μ^*) is a sequential equilibrium.

Consider next the following pair of properties that a behavioral strategy may satisfy:

(S^*2) Any player who receives a message in M reveals that message.

(S^*5) Any player who does not receive a message in M quits when he is chosen to be either P or FR. If a certified player announces that he agrees to match with a player who has no message, then this player accepts such an offer.

Consider any strategy satisfying ($S1$), ($S3$), and ($S4$), and either ($S2$) and ($S4$), or (S^*2) and (S^*4). Denote such a strategy profile by β^{**}. Let μ^{**} be a belief system derived from β^{**} using Bayes rule, which also satisfies $(B1) - (B5)$. Then using essentially the same arguments as above one can show that (β^{**}, μ^{**}) is also a sequential equilibrium of $\Gamma\left(G^{SU}\right)$. Note that in both (β^{**}, μ^{**}) and (β^*, μ^*) all players test, no infections occur and the only players to match are healthy with healthy and infected with infected.

Step 2: *Uniqueness of equilibrium outcome.*

Assume that (β', μ') is a sequential equilibrium of $\Gamma\left(G^{SU}\right)$ differing from the two types of equilibria described in Step 1. Suppose that all players test in (β', μ'). For μ' to be consistent, the interpretations given to each message (including the null one) must be the same as in the above equilibria.

Suppose that a nonempty subset of players reveal their certificates whenever they are certified. Denote this subset by R. Assume player 2 does not belong to R. Then by revealing his certificate, a certified player 2 strictly increases his chances of a match with any player in R who is certified. To see why, note first that when a player in R is certified and selected to make a proposal (as either P or FR), he will choose C (since a certified player will not match with anyone who is *not* certified, he does not care whether uncertified players choose C as P or FR). Second, consider a history in which player 2 was selected as P, or one in which he was selected to make a proposal

as FR and SR is a certified player from R. If player 2 chooses C, a certified player from R will surely accept. In every other history player 2's expected payoff equals his payoff before the deviation. It follows that there exists no sequential equilibrium in which only one or two players reveal their certificates.

Suppose no player reveals himself when certified. Then each player is believed to be infected with probability p and players would agree to match with one another. Thus each player has a probability of $2/3$ to match. By sequential rationality, after any history in which a certified player reveals himself, any other player who is chosen to be P will choose to match only with a certified player (who, by sequential rationality, will agree to match with a player who is infected with probability p). Thus, any player who is certified will want to deviate and reveal himself, a contradiction.

From the argument above it follows that it cannot be that all players test in (β', μ'). Suppose no player tests in (β', μ'). By design, when the two other players do not test, the expected payoff of a player who tests and is certified (resp., not certified) in $\Gamma\left(G^{SU}\right)$ is equal to the payoff of a player who tests in $\Gamma(G)$ and receives a message in M_1 (resp., M_2). In addition, the payoff of a healthy (resp., unhealthy) player from matching with an untested individual is the same in both $\Gamma\left(G^{SU}\right)$ and $\Gamma(G)$. By the convexity of A, the anxiety of an unmatched player who does not receive a message in $\Gamma\left(G^{SU}\right)$ cannot be higher than the average anxiety of an unmatched type 3 player (where the average is taken over all messages in M_3) in the game $\Gamma(G)$. Therefore, if there is a no-testing equilibrium in $\Gamma\left(G^{SU}\right)$, then there must be one in $\Gamma(G)$, a contradiction.

Suppose (β', μ') is a sequential equilibrium in which one or two players test. Consider the lottery that a player faces when he tests in the game $\Gamma\left(G^{SU}\right)$: With probability x the tester will be certified as healthy for sure, with probability y he will be certified as infected for sure and with probability $1-x-y$ he will know that he has a $\frac{p-y}{1-x-y}$ chance of being infected. Now consider a player who tests in the game $\Gamma(G)$: x is the probability of receiving a message in M_1, y the probability of receiving a message in M_2 and $\frac{p-y}{1-x-y}$ the probability of receiving a message in M_3. By the convexity of A, for a given match (including no match), the anxiety of a player who does not receive a message in $\Gamma\left(G^{SU}\right)$ is smaller than the expected anxiety of a player who receives a message $m \in M_3$ in the game $\Gamma(G)$ (where the expectation is taken over all the possible messages in M_3). This means that if there is a sequential equilibrium in $\Gamma\left(G^{SU}\right)$ in which at least one player does not match, then there must also be an equilibrium with that property in $\Gamma(G)$, a contradiction. This completes the proof of Claim 1.□

By Claim 1, the game $\Gamma\left(G^{SU}\right)$ has only two sequential equilibria. Both equilibria achieve zero infections and the same volume of trade as in $\Gamma(G)$ (in both games the probability of a match is $x+y$). However, the anxiety of a player who receives no message in the two equilibria of $\Gamma\left(G^{SU}\right)$ is never higher than the average anxiety of all type 3 players in $\Gamma(G)$. Thus, according to the preferences of the planner, the unique equilibrium outcome of $\Gamma\left(G^{SU}\right)$ is not inferior to that of $\Gamma(G)$.■

6 Infection-Free Unconditional Mechanisms

The last result greatly simplifies our search for an unconditional mechanism implementing zero infections. The proof establishes that such a mechanism, if it exists, contains two messages that are highly specific. If an individual tests positive, then he is either given no message, or a certificate implying that he is infected for sure. If an individual tests negative, he is either given no message, or a certificate implying that he is certainly not infected.

It turns out that there is additional power in assumption A2 that further simplifies our search. With A2, it can be established that there is a very simple (one message) mechanism that implements zero infections. This message is the clean bill of health given to some proportion of those who are in fact healthy. We refer to mechanisms of this type as certification schemes, since they are so strongly analogous to the certification policies discussed by P&P.

Definition 6 *A certification policy is a very simple unconditional mechanism that never certifies an individual who tests positive:* $M = \{C\}$*, and* $t_i = + \Longrightarrow g_i(\emptyset \mid t) = 1$

Proposition 4 *Given A1 and A2, there exists a certification scheme that implements zero infections.*

P roof. Certification schemes can be fully characterized by the level of belief $b \in (p, 1]$ about the probability of sickness given that one has not been given a certificate. Note that from an ex ante perspective, it must be that a proportion $\frac{p}{b}$ of the population expect to not receive certificates, with the remainder being certified. Consider a specific certification scheme in which the ex post probability of sickness for an uncertified individual is precisely $b = \frac{2}{H} < \frac{1}{2}$. (Note that $p < \frac{2}{H}$, so this is definitely a feasible mechanism). We first show that such a mechanism produces an equilibrium in which all indeed test, and only those who pass end up matching. Simple symmetric strategies delivering that outcome are as follows.

In the first stage of the game all players test. A player who is certified reveals his certificate. The strategy of each player at each information set in the matching subgame are as follows:

- If selected as P choose C if there is one, otherwise quit.
- If selected as FR accept an offer from a certified P, reject otherwise. If P quits, or if FR rejects P's offer, then make an offer to SR only if SR is certidied.
- If selected as SR, accept an offer only if FR is certified.

Beliefs are updated according to the strategies. In particular, at any information set a certified player is believed to be healthy while a noncertified player is believed to be infected with probability $\frac{2}{H}$.

The fact that certified individuals are acceptable as partners is an obvious requirement of sequential rationality. What needs explaining is why non-certified individuals, untested individuals, and certified individuals all reject any matches with those who are uncertified. The key here is assumption A2, which makes such a match too costly, purely on health grounds. A2

is designed to ensure that the expected health loss from such a match between two uncertified individuals exceeds the matching benefit,

$$H(1-\frac{2}{H})\frac{2}{H} = 2 - \frac{4}{H} > 1.$$

Combining A1 and A2, it is clear that the corresponding inequality holds for an untested individual contemplating a match with an individual believed to have failed the test, for whom the health cost is:

$$H(1-p)\frac{2}{H} = 2(1-p) > 2(1-\frac{1}{H}) > \frac{3}{2}.$$

Finally, it is immediate that certified individual will reject a match with an individual believed to have failed the test, since here the net loss from a match on pure health grounds is precisely 1.

Using essentially the same arguments as in the proof of Proposition 1, it is straightforward to verify the consistency of the above assessment. The consistency of the belief system requires each player to believe that a certified player is healthy for sure. This means that a certified player will always be offered a match, and no player would decline an offer to match with a certified player. Since a match with another certified player gives a net benefit of 1, revealing one's certificate is sequentially rational.

It follows that in order to confirm that the assessment we proposed is indeed an equilibrium, it remains only to show that it is worthwhile for one individual to test if the others are known to be testing. The only trade off is matching benefits against anxiety costs, since there is no additional infection risk either from testing or from defecting to non-testing. With respect to matching costs, it is clear that not testing results in no matches taking place, while testing and becoming certified yields a match in two circumstances: a certainty of matching if one and only one of the other two pass, and a $\frac{2}{3}$ chance of matching if the other two do. On the other hand, testing yields an increase in anxiety when the test is failed. For a general test of the type above with ex post belief $b \in (p, 1)$ conditional on receipt of a bad signal, the incremental anxiety from testing when the other two test is $\frac{pK(b-p)^r}{b(1-p)^r}$: if the test is failed (probability $\frac{p}{b}$), then the ex-post belief is b. The gain in match utility from taking the test is $\frac{2}{3}(1-\frac{p}{b})^3 + 2\frac{p}{b}(1-\frac{p}{b})^2$, corresponding to the $\frac{2}{3}$ chance of matching if all three pass and the definite match as one of only two who certify.

Before confirming that A2 implies that the benefits outweigh the losses, we first note a general upper bound on the anxiety cost valid for all $b > p > 0$:

$$\frac{pK(b-p)^r}{b(1-p)^r} < pKb^{r-1}.$$

With this we can write the general condition for the anxiety cost to be no higher than the trade benefit as,

$$pKb^{r-1} \leq \frac{2}{3}(1-\frac{p}{b})^3 + 2\frac{p}{b}(1-\frac{p}{b})^2.$$

Substitution of $b = \frac{2}{H}$ for the test at hand yields,

$$\left(\frac{2}{H}\right)^{r-1} \leq \frac{2(1-\frac{pH}{2})^3 + 6\frac{pH}{2}(1-\frac{pH}{2})^2}{3pK}.$$

The truth of this inequality follows from A2,

$$2(1-\frac{pH}{2})^3 + 6\frac{pH}{2}(1-\frac{pH}{2})^2 > (1-pH)^2$$

We now turn to show that there exists no (pure strategy) equilibrium, other than the one we have described above. First we show that there exists no sequential equilibrium in which at least one person tests, and some certified player does not reveal his certificate. Assume there exists a sequential equilibrium in which at least one player reveals his certificate when he receives one, but that some other player does not. Assume w.l.o.g. that player 1 does not reveal himself when certified. We need to show that by revealing himself, player 1 would strictly increase his expected payoffs. There are two cases to consider.

Case 1: *A certified player will not match with a player who does not show a certificate.* The set of all possible histories can be partitoned into two sets Ω_1 and Ω_2. The first set Ω_1 contains four types of histories: (1) Player 1 is certified, he is chosen as P and one of the two other players is certified and has revealed his certificate, (2) player 1 is certified, he is chosen as FR and SR is certified and has revealed himself, (3) P is a certified player who revealed himself and (4) player 1 is SR and FR is a certified player who revealed himself. The set Ω_2 contains all the remaining histories. In the proposed equilibrium (in which player 1 does not reveal), in all the histories in Ω_1 player 1 is not matched with the certified player. However, when player 1 deviates and reveals his certificate, he is sure to match with the certified player in every history in Ω_1. In all the histories in Ω_2, player 1's expected payoffs are the same before and after his deviation. It follows that player 1 would gain by deviating from the proposed equilibrium, a contradiction.

Case 2: *A certified player will match with a player who does not show a certificate.* In this case by deviating and revealing his certificate, player 1 would strictly increase the probability of matching in every history, a contradiction.

From the analysis of the above two cases we conclude that there exists no sequential equilibrium in which only one or two players reveal themselves when they are certified. We now show that there exists no sequential equilibrium in which no player reveals a certificate.

Suppose there exists a sequential equilibrium in which no player who is certified reveals his certificate. Because the mechanism is unconditional, each player is believed to be infected with probability p (as if no player has tested). By A1, players would want to mtch with one another. Since all players are identical in the matching game, each is matched with probability 2/3. Thus, any certified player who reveals himself will increase his chances of matching to one. It follows that there cannot be a sequential equilibrium with no revelation.

A second point is that regardless of others test strategies, the anxiety loss following from a decision to test rather than to not test is always below $pK\left(\frac{2}{H}\right)^{r-1}$. To see why, note that by (A2), there cannot be an equilibrium in which there is some history after which a player agrees

to match with a noncertified player who tested. Therefore, the maximal anxiety level that can be reached in equilibrium occurs when a player takes a test, is not certified and then matches for sure with an untested player. This anxiety level is equal to $\frac{pK(b-pb)^r}{b(1-p)^r} = pK\left(\frac{2}{H}\right)^{r-1}$. Hence, one can guarantee that in equilibrium each player strictly prefers to test provided the classical net gains from trade (matching and health only) exceed this bound. What this means in light of (A2) is that it suffices to show that these benefits exceed $\frac{(1-pH)^2}{3}$.

Consider first a potential equilibrium in which the two others test, but the one we are studying is presumed not to test. For simplicity, say that it is 2 and 3 who test. Then the equilibrium strategies of players 2 and 3 must satisfy that whenever they are certified, they reveal their certificate if they are chosen to make an offer. What benefits occur to 1 if he tests and passes rather than not test? First, given the other players' revelation strtaegies, player 1 would reveal his certificate. Second, if both others pass, he gets a $\frac{2}{3}$ probability of matching rather than no chance. If one other passes, he gets a 100% chance as opposed to none. The reason player 1 would not match, if he does not test, is as follows. By A3 ($H > 4$), even a $\frac{1}{2}$ probability of matching with a taker and failer of the test is no good on pure health grounds for the one who is clear, since the resulting health loss is,

$$H(\frac{p+b}{2}) > \frac{bH}{2} = 1.$$

In addition, an untested player will not want to match with a player who took the test and failed. This means that the trade gain in this case is $\frac{2}{3}(1-\frac{p}{b})^3 + 2\frac{p}{b}(1-\frac{p}{b})^2$, which certainly exceeds $\frac{(1-pH)^2}{3}$.

If only player 2 is supposed to take the test and passes, then player 1 who is not supposed to take the test certainly benefits from testing, passing and revealing his certificate: in this situation he is the partner for sure as opposed to with only a $\frac{1}{2}$ chance. The matching gain is,

$$\frac{1}{2}(1-\frac{p}{b})^2 = \frac{1}{2}(1-\frac{pH}{2})^2.$$

This again exceeds $\frac{(1-pH)^2}{3}$ ■

This last result characterizes one mechanism that implements zero infections. Given that such a mechanism exists, and given the planner's lexicographic preferences, we know that her optimal (or ε-optimal) mechanism will therefore be of this variety. In fact, it is clear from the proof of Proposition 4 that there is a connected set of certification schemes with varying levels of ex-post belief that satisfy all of the required conditions. There are other such mechanisms involving the sure infection message for some proportion of those who test positive. Among all of these the planner will select based on her secondary goal of maximizing matching. Of course, she will be unable to guarantee matches 100% of the time, since this would require her to use a fully-revealing test, in which case Proposition 2 has already established that there is a no testing equilibrium.

7 Infection-Free Conditional Mechanisms

When we turn to conditional mechanisms, the planner will be able not only to keep infections to zero, but also to generate full trade. Furthermore she will be able to keep anxiety at its absolute minimum consistent with these two outcomes. This is absolutely her first best outcome.

Definition 7 *The **first best** outcome satisfies the following:*
(1) *The probability that a player will infect another is zero.*
(2) *The probability that two players will match is one.*
(3) *Any outcome in which the expected anxiety is strictly lower violates either* (1) *or* (2)*.*

Our proof that this outcome can be implemented using a conditional mechanism is constructive, and is based on a specific mechanism, which we call the Minimally Informative Guidance mechanism (MIG).

Definition 8 *The **MIG** mechanism is a very simple conditional mechanism with the following properties:*
(1) $M = \{C\}$
(2) *If all test, then only the two players with identical test results (if all three have identical test results, then one pair is randomly chosen) receive the message C.*
(3) *If not all test, then an unconditional mechanism is carried out: With probability ε each player who tests is given C if healthy, no message if sick. With probability $1-\varepsilon$ each player who tests receives the message C regardless of his test result.*

Proposition 5 *Under assumptions A1 and A2, the MIG mechanism implements the first best outcome: the unique equilibrium outcome of $\Gamma(MIG)$ is the first best.*

P roof. The proof proceeds as follows. First, we show that there exists an equilibrium in $\Gamma(MIG)$ in which no infections occur and matching always takes place. We then show that the expected anxiety of each player is the minimum possible subject to having no infections and sure matches. Finally, we prove uniqueness.

Step 1: There exists an equilibrium with no infections and efficient matching.
Let (β, μ) be an assessment with the following properties. In the first stage of the game all players test. Any player who is certified reveals his certificate. The strategies in the matching subgame depend on the history that led to that subgame. In particular, each player conditions his matching strategy on whether or not he tested, on the revealed profile of messages and on his role in the matching game (whether he is selected as P, FR or SR).

- Consider the matching subgames that follow a history in which two players revealed their certificate. The players' behavioral strategies are as follows:
  - If selected as P, choose C.
  - If selected as FR when P does not quit, accept an offer made by a certified player.
  - If selected as FR when P quits, make an offer to SR only if he is certified, otherwise, quit.

– If selected as SR accept only offers made by a certified player.

Note that according to the above strategies an uncertified player will match only with a certified one, since these strategies apply only to information sets in which there is at most one uncertified player.

- Consider the matching subgames that follow a history in which exactly one player revealed his certificate. The players' behavioral strategies are as follows:

  – The certified player quits if selected to make a proposal (either as P or as FR). As a responder, the certified player will decline all offers.

  – An untested player will agree to match only with the certified player. Therefore, when making a proposal (as either P or FR) he chooses C if the certified player had not quit, otherwise he himself quits. When responding to an offer (as either FR or SR), he accepts only if made by the certified player.

  – A tested an uncertified player is infected for sure. He would therefore like to match with any player. As a proposer, he either chooses C, $\emptyset$, $\{C \cup \emptyset\}$ or quits. As a responder, he accepts any offer, which is not made exclusively to certified players.

- Finally, consider those matching subgames that follow a history in which no player revealed his certificate. The players' behavioral strategies are as follows:

  – An untested player will not want to match. Therefore he will quit as a proposer and decline all offers.

  – A tested but uncertified player is infected for sure. Therefore, as described for informations sets with only one certified player, there are several strategies he might use when proposing. As a proposer, he either chooses C, $\emptyset$, $\{C \cup \emptyset\}$ or quits. As a responder, he accepts any offer, which is not made exclusively to certified players.

Beliefs are updated according to the strategies. In particular, at any information set in which at least two individuals are certified, each player holds the following beliefs: a certified player is believed to be infected with probability b_C, while a noncertified player is believed to be infected with probability $b_\emptyset$ where:

$$\begin{aligned} b_C &= 3p^2(1-p) + p^3 < p \\ b_\emptyset &= 3[p - 2p^2(1-p) - \frac{2}{3}p^3] = 3p - 6p^2 - 4p^3 \end{aligned}$$

Any information set in which less than two players were certified cannot be reached if players follow the precepts of β. This means that the beliefs of the players at those information sets cannot be derived from the strategies using Bayes rule. Therefore, we need to explicitly specify the players' out-of-equilibrium beliefs.

At any information set in which exactly one player was certified the players hold the following beliefs. First, it is common knowledge among the players that if only one player was certified,

then at least one player *did not test*. This means that the beliefs of the players will depend on their message (whether or not they were certified) and on their own past action (whether or not they have tested). The certified player believes that exactly one of the two other players did not test. He therefore believes that a noncertified player is infected with probability $\frac{1}{2}(1+p)$. Both an untested player and a noncertified player who did test believe that a certified player is healthy with the following probability

$$\frac{(1-p)}{1-\varepsilon p}$$

However, the noncertified untested and the noncertified tested players disagree on the interpretation they give to a noncertification. An untested player believes that the noncertified player is an individual who tested and failed and is, therefore, infected with probability 1. A noncertified player who tested knows that at least one player did not show up for the test. Since he knows that he himself tested and that the certified player certainly tested, he infers that it must be the other noncertified player who did not test. Therefore, a noncertified but tested player believes that another noncertified player is infected with probability p.

At any information set in which no player was certified players hold the following beliefs. As in information sets with only one certificate, it is common knowledge among the players that at least one player did not test. Any player who did not test believes that a noncertified player is a player who tested and failed and is therefore infected for sure. Any player who tested but failed believes that exactly one of the other noncertified players did not test. He therefore believes that a noncertified player is infected with probability $\frac{1}{2}(1+p)$.

We now verify that (β, μ) is an equilibrium. We begin by verifying the sequential rationality of the strategies in the matching subgames. Given μ and given that certified players reveal themselves, it is clear that the players' strategies are optimal in any matching subgame that follows a history in which at least two players are certified. Consider next all the subgames that follow a history in which at most one player was certified. It is straightforward to verify that given μ, the strategies of the noncertified players (the ones who have tested and the ones who have not) are optimal. Given μ, there exists a sufficiently small ε such that a certified player will not want to match with a noncertified player if the following inequality holds:

$$1-\left(\frac{1+p}{2}\right)H-\frac{K}{2^r}<0 \tag{1}$$

A sufficient condition for (1) to hold is $H>4$, which is assumed in $(A2)$.

We now turn to the revelation stage. Revealing one's certificate is weakly dominating in any subgame that follows a history in which a player was certified. Hence, it is sequentially rational to reveal a certificate whenever one is given.

Finally, we turn to show that no player can gain by deciding not to test. When all test each player is matched with probability $\frac{2}{3}$ and his net value from matching is 1 (there are no infections). The expected anxiety of each player is simply $\frac{1}{3}A(b_\emptyset)$ (the probability of being certified is $\frac{2}{3}$ and $A(b_C)=0$ since $b_C<p$).

Suppose one player, say j, deviates from β by not testing. Then j will receive the null message with certainty and will remain unmatched. It follows that his payoff from deviating is.

$-A(p) = 0$. Thus, to verify that j cannot gain by deviating it suffices to show that when player j does not deviate from β his expected utility is strictly positive, i.e.:

$$\frac{2}{3} - \frac{1}{3}A\left(b_{\emptyset}\right) > 0 \tag{2}$$

where

$$A\left(b_{\emptyset}\right) = \frac{K(2p - 6p^2 - 4p^3)^r}{(1-p)^r} \tag{3}$$

By substituting 3 into 2, the latter inequality can be rewritten as follows:

$$2^{r-1}K < \frac{(1-p)^r}{(p - 3p^2 - 2p^3)^r} \tag{4}$$

Similarly, we rewrite $(A2)$ as follows:

$$2^{r-1}K \leq \frac{H^{r-1}(1-pH)^2}{3p} \tag{5}$$

Thus, in order to verify that 4 holds, it suffices to show that the RHS of that inequality is strictly larger than the RHS of 5, i.e. we need to show that

$$\frac{(1-p)^r}{(p - 3p^2 - 2p^3)^r} > \frac{H^{r-1}(1-pH)^2}{3p} \tag{6}$$

A sufficient condition for 6 is simply $(A2)$. This establishes that β is sequentially rational.

To complete Step 1 of the proof it remains to show that (β, μ) is consistent. We construct the following pertubation of (β, μ), which we denote by (β^ρ, μ^ρ). At each information set I_i the behavioral strategy $\beta^\rho\left(I_i\right)$ assigns a probability of ρ to any action available at I_i and which was not assigned a positive probability by $\beta\left(I_i\right)$. The remaining probability is assigned to the action chosen by $\beta\left(I_i\right)$. The belief system is then derived using Bayes rule. Clearly, as $\rho \to 0$ we have that $\beta^\rho \to \beta$. Also, the beliefs assigned by μ^ρ to information sets, that are reached when all players test, converge to the beliefs assigned by μ to those information sets as $\rho \to 0$. It remains to show that the beliefs assigned by μ^ρ to information sets, that are reached when at least one player did not test, converge to the beliefs assigned by μ to those information sets as $\rho \to 0$.

Any information set in which exactly one player was certified can be reached if and only if one of the following occured: (1) Only one player tested or (2) only two players tested, one of which was not certified. By the construction of β^ρ, as $\rho \to 0$ the probability that two players did not test approaches zero. This implies that as $\rho \to 0$, the probability that two players have not tested, conditional on them being noncertified, approaches zero. Therefore, the conditional probability, that only one of those two noncertified players did not test, goes to one.

Any information set in which no player was certified can be reached if and only if one of the following occured: (1) Only one player tested and he failed, (2) only two players tested and both failed or (3) no player tested. A player who tested does not know whether (1) or (2) occured. As $\rho \to 0$ the probability of (1), conditional on either (1) or (2) occuring, goes to zero. Hence, the conditional probability of (2) goes to one. This proves that the beliefs of a tested player at

information sets, in which no player was certified, is consistent. An untested player does not know whether (1), (2) or (3) occured. Similarly, as as $\rho \to 0$, the conditional probability of (2) goes to one. This proves that the out-of-equilibrium beliefs of untested players are consistent. It follows that $(\beta^\rho, \mu^\rho) \to (\beta, \mu)$ as $\rho \to 0$, which completes the proof of consistency.

Step 2: The equilibrium outcome of Step 1 is unique.

Let $\left(\beta^3, \mu^3\right)$ be a sequential equilibrium in which all test. Consistency requires each player to believe that a certified player is healthy with probability b_C. Hence, each player's most preferred outcome in the matching subgame is to be matched with a certified player. By sequential rationality, every equilibrium must satisfy that each player agrees to match with a certified player and each player who proposes a match will propose to a certified player, if one exists. Thus, if all players reveal their certificates, the only equilibria in which all test are those described in Step 1.

Assume there exists a sequential equilibrium in which all test, but only one player reveals his certificate. W.l.o.g. assume that players 1 and 2 do not show their certificates. We claim that one of these players, say 1, can strictly increase his expected payoff by deviating and revealing his certificate (whenever he receives one). To prove our claim it suffices to show that by revealing himself player 1 increases his payoff in all the histories in which he is certified. Since all players test, then whenever player 1 is certified there is exactly one other player who is certified as well. There are two possible contingencies we need to consider.

1. *Player* 2 *is certified.* When this contingency occurs each player would find himself at at information set in which no one has revealed a certificate. The only history which can lead to those information sets and which is consistent with the equilibrium strategies, is one in which all tested and players 1 and 2 were certified. That history must be assign a probability of one in order for the players' beliefs to be consistent. Thus, players 1 and 2 would like to match with one another. The only history in which player 1's decision whether or not to reveal his certificate would make a difference is one in which player 2 is chosen to be P. When player 1 does not reveal his certificate player 2 can either quit or make an offer to both players 1 and 3. Thus, the chances that player 1 would end up matching with player 2 are at most 50%. However, by revealing himself, player 1 would allow player 2 to choose C and would therefore raise his chances of matching with player 2 to a 100%.

2. *Player* 3 *is certified.* Player 1's most preferred outcome in this matching subgame is to match with player 3. However, if player 1 does not reveal his certificate, then player 3 cannot make an exclusive offer to player 1 when 3 is chosen as P. Thus, the probability that player 1 and 3 would match when 3 is chosen as P is at most a half. However, if player 1 would deviate from his non-revelation strategy, then he can increase his chances of matching with player 3, whenever 3 is selected as P, to one. In all other histories, player 1 gets the same payoff whether or not he reveals his certificate.

From the analysis above, it follows that player 1 can strictly increase his expected payoff by revealing his certificate. Hence, there cannot be an equilibrium in which only two players reveal themselves.

Assume there exists an equilibrium in which all test, but no player reveals his certificate. The only consistent belief a certified player may hold is that all have tested and one of the other two players is certified. From the same argument used in Cases 1 and 2 above, it follows that each player who is certified can strictly increase his expected payoff by revealing his certificate, a contradiction.

Finally, assume there exists an equilibrium in which all test, but only two players reveal themselves. W.l.o.g. let player 1 be the player who does not reveal himself whenever he is certified. The only histories that lead to outcomes, which are different from those obtained in the equilibria of Step 1, are those in which player 1 is certified. Let Ω denote that set of histories. Consider a matching subgame that follows a history in Ω. Assume w.l.o.g. that player 2 was also certified (since all test and 1 was certified, one of the two other players must have been also certified). Each player then observed that only player 2 revealed his certificate. From the knowledge of the equilibrium strategies players 2 and 1 infer that player 3 tested but was not certified. Hence, players 1 and 2 would agree to match only with one another.

We now show that player 1 can increase his expected payoff by revealing his certificate. The matching game that follows the revelation stage starts with one of the players being chosen to be P. Suppose player 2 is selected to be P. Then he can either quit or choose $\emptyset$. If player 2 does make a proposal, he must accept a 50-50 chance of matching with either a certified or an uncertified player. This also means that player 1 has at most a 50-50 chance of matching with player 2. Player 1 can increase that probability to one by revealing his certicate before the matching subgame begins. Suppose player 1 or 3 are selected as P. Then player 1's payoff remains the same whether or not he reveals his certificate. It follows that revealing a certificate is a profitable deviation for player 1, in contradiction to our initial assumption.

Assume there exists an equilibrium $\left(\beta^{\emptyset}, \mu^{\emptyset}\right)$ in which no player tests. By consistency, each player believes that a certified player is a better match than a noncertified player. By $(A1)$, an untested player and a healthy players both agree to match with an untested player. By sequential rationality, if one of the other players is known to be certified, then a player in the role of P will offer a match to that certified player. Similarly, each player will accepts an offer to match with a certified player.

Consider a deviation from $\left(\beta^{\emptyset}, \mu^{\emptyset}\right)$ by one of the players, say j, such that j tests in the first period, and reveals his certificate. From the arguments presented above, it follows that this deviation is strictly profitable to player j. To see why, note that by design player j is certified with probability $1 - p\varepsilon$. This means that j will match with that same probability. The health costs incurred by player j can be of two types: (1) The same costs he faced when he matched as an untested individual ($(1-p)pH$), or (2) the costs of a healthy individual who is matched with an untested individual ($(1-p)H$). The probability that j will incur the first type of health costs, given that he matched, is $1-\varepsilon$, while the probability that he will incur the second type of health costs is $\varepsilon(1-p)$. The final component of j's expected utility after testing is his anxiety. The only circumstance in which j's anxiety when he tests will be different from his anxiety when he does not test, is when G runs a truthful test, the probablity of which is ε. Thus, there exists an ε sufficiently close to zero for which player j's deviation is strictly profitable, a contradiction.

Assume there exists an equilibrium $\left(\beta^1, \mu^1\right)$ in which exactly one player tests. In this equilibrium each of the two untested players stand an equal chance of being matched. Suppose one of the untested players, say j, deviates from his equilibrium strategy by showing up for the test and revealing his certificate. Then whatever μ^1 is, the probability that j will match is at least $1-\varepsilon(1-p)$. Moreover, player j's expected health costs of matching will decrease since he will match with a certified player with probability $(1-\varepsilon)(1+\varepsilon(1-p))$. There are only two circumstances in which player j's anxiety will go up: (1) When he is not certified, or (2) When he is certified, but faces two uncertified individuals. The probability of (1) and (2) occuring is εp and $\varepsilon p[2-\varepsilon p]$ respectively. In all other circumstances, player j's anxiety will decrease as a result of his deviation. Thus, there exists an ε sufficiently close to zero for which player j's deviation is strictly profitable, a contradiction.

Finally, assume there exists an equilibrium in exactly two players test. Then the untested player (say, j) almost surely never matches. By deviating to testing and showing his certificate, player j is guaranteed a match with probablity $\frac{2}{3}$. However, as a result of his deviation, player j will have a strictly positive level of anxiety with probability $\frac{1}{3}$. As shown in Step 1 of the proof, when all three players test the net (physical) benefits from matching is strictly greatr than the cost of anxiety. It follows that it is strictly profitable for player j to deviate and test, a contradiction. This completes the proof of Step 2.

Step 3: Minimal anxiety.

A necessary condition to have a match in every contingency and to eliminate the risk of infections is the following: In every contingency two players of the same health status are matched. Since $p < \frac{1}{2}$ any unmatched player necessarily infers that he is more likely to be infected than the players who are matched. The expected probability that an unmatched player is infected is precisely $b_\emptyset$. Similarly, a match player necessarily infers that being matched does not guarantee that one is healthy. Thus, the expected probability that an unmatched player is infected is precisely b_C. The only thing that a more elaborate mechanism could do would be to provide additional information. However, if one provides more information, all it can do is to perform mean preserving spreads around these beliefs, which would necessarily raise anxiety in light of Jensen's inequality. ■

8 Credibility

In the previous sections we proposed two mechanisms for implementing the planner's objective. We implicitly assumed that the planner could commit to these mechanisms. However, if the planner actually participates in the mechanism by being in charge of sending the messages to the agents, then she becomes a player in the game just like the agents. In such a case, it may not be reasonable to assume that the planner can commit to any strategy. For example, suppose the planner is a physician whose strategy is not to report truthfully a negative result. Then she may be tempted to stray from this strategy when faced with an anxious patient who tested negative.

To study the question of credibility we analyze the policy maker's problem when she actually plays in the game that is induced by the policy that she designs. Problems of this type were first analyzed by Baliga, Corchon and Sjöström (97) (henceforth, BCS) who introduced the theory of

implementation when the planner is a player. As demonstrated by BCS, in a situation where the planner cannot commit to a mechanism and the outcome function is substituted by the planner himself, the implementation problem is transformed into a signaling game.[8]

Let Γ^* be the following extensive form game with imperfect information. In the first period the three players simultaneously decide whether or not to test. If at least one player decides to test, then after seeing the test results the planner chooses a lottery g over M, subject to the constraint that an untested individual receives no message. Following the planner's decision and after the lottery that he chose is carried out, the players engage in the same matching game that we described in the previous sections.

Recall that given a pair of outcomes with identical rates of infections and matching, the planner prefers the outcome in which the average level of anxiety is lower. Hence, the planner's payoff is affected by the players' *beliefs*. This transforms the signaling game Γ^* into a *psychological game*. Psychological games were introduced by Geanakopolos, Pearce and Stacchetti (1989) (henceforth GPS). Unlike standard games, in psychological games the payoff to each player depends not only on what every player does, but also on what he thinks every player believes. In order to analyze this new class of games, GPS introduced the notion of a *psychological equilibrium (PE)*. This new notion of equilibrium consists of a profile of strategies and beliefs such that (1) given the players' beliefs, the strategy of each player is a best response to both the strategies of the other players *and their beliefs* (2) each player's belief is consistent with the equilibrium strategies.

Our objective in this section is to investigate the credibility of the mechanisms that we proposed in Sections 4 and 5.

Definition 9 *Let $G \in \left\{\mathcal{G}^U, \mathcal{G}^C\right\}$ be a mechanism that implements the planner's social welfare function. G is said to be* ***credible*** *if the game Γ^* has a sequential equilibrium in which the following is satisfied:*
(1) *The outcome of that equilibrium is exactly the equilibrium outcome of $\Gamma(G)$.*
(2) *The planner's equilibrium strategy is exactly what G prescribes.*

Note that for implementation we insisted on a unique desirable equilibrium, whereas for crediblity we only require that at least one desirable equilibrium exists. The reason is that when the planner enters the game as a player the original game is transformed into a game of signalling, which typically has multiple equilibria. Some of equilibria can be eliminated by positing additional restrictions on the out-of-equilibrium beliefs, discussion of which will obscure our point.

Proposition 6 *The optimal unconditional mechanism is incredible.*

P roof. Assume that the optimal unconditional mechanism is credible. Then Γ^* has a sequential equilibrium with reasonable beliefs that satisfies the following properties:

[8] Directly related to the current game is an example in Caplin and Leahy [1999] in which a planner who cares about private welfare is unable to resist passing on good news, with the unfortunate side effect that no news is bad news.

1. All players test.

2. If all players pass the test, the probability that all are certified is strictly less than one.

3. Each player believes that a certified player is healthy for sure.

4. Each player would like to match with another certified player.

Consider a history in which all test and pass. The planner's best response is to certify all three, a contradiction. ■

Proposition 7 *The MIG mechanism is incredible.*

P roof. Assume the MIG mechanism is credible. Then Γ^* has a sequential equilibrium (β, μ) with the following property: If only a single player shows up for the test, then the planner uses a mixed strategy in which with probability ε he certifies the player only if the player tested negative, and with probability $1 - \varepsilon$ he certifies the player regardless of the player test results. We will now show that this behavioral strategy is not sequentially rational.

Consider the subgame that follows the history in which only player 1 decided to test. For μ to be consistent, players 2 and 3 must believe that if player 1 is non-certified, then he must be infected for sure. Thus, if player 1 tests positive, then the unique optimal strategy for the planner is not to certify the player with certainty. Thus, β does not satisfy the property described above, a contradiction. ■

These negative results on credibility provide a cautionary note on our earlier positive results on implementation. Of course the delicate question is the extent to which the credibility constraints matter. For the unconditional mechanism, it would be relatively easy to mechanize the output of the test, thereby removing all choices. In fact, one can read the proposition as pointing out the directions in which to improve actual AIDS tests: it may be more important to accurately rule out the disease than to positively identify it. For the conditional mechanism, the machine that would be needed to carry out the optimal strategy would be somewhat complex, and could not conceivably be viewed as simply an imperfect AIDS test, since it sometimes certifies people who are known to be infected. Of course, if such mechanisms are viewed a priori as absurd, then we should be limiting our search to some subset of the unconditional mechanisms that are less flawed. Rather than trying to impose such constraints from the outset, we have chosen the alternative strategy of fully exploring the unconstrained setting. We see this as a necessary prelude to more realistic modeling.

References

[1] Baliga, S., L. Corchon and T. Sjostrom, "The Theory of Implementation When the Planner is a Player," *Journal of Economic Theory* **77** (1997), 13-33.

[2] Caplin, A. and J. Leahy, "The Supply of Information by a Concerned Expert," Research Report **99-08** (1999), C.V.Starr Center, New York University.

[3] Caplin, A. and J. Leahy, "Psychological Expected Utility and Anticipatory Feelings," *Quarterly Journal of Economics* **116** (2001), 55-80.

[4] Geanakoplos, J., D. Pearce, and E. Stacchetti, "Psychological Games and Sequential Rationality," *Games and Economic Behavior* **1** (1989) 1, 60-79.

[5] Jackson, M.O., "A Crash Course in Implementation Theory," *Social Choice and Welfare* **18** (2001) 4, 655-708.

[6] Kremer, M, "Integrating Behavioral Choice into Epidemiological Models of AIDS," *Quarterly Journal of Economics* **110** (1996), 549-573.

[7] Kreps, D., and E. Porteus, "Temporal Resolution of Uncertainty and Dynamic Choice Theory", *Econometrica* **46** (1978), 185-200.

[8] Kreps, D., and R. Wilson, "Sequential Equilibria", *Econometrica*, **50** (1982), 253-279.

[9] Lyter, D., R. Valdiserri, L. Kingsley, W. Amoroso, and C. Rinaldo, "The HIV Antibody Test: Why Gay and Bisexual Men Want or Do Not Want To Know Their Results", *Public Health Reports* **102** (1987), 468-474.

[10] Osborne, M., and A. Rubinstein, *A Course in Game Theory* (1994), MIT Press, Cambridge, MA.

[11] Philipson, T. and R. A. Posner, "A Theoretical ans Empirical Investigation of the Effects of Public Health Subsidies," *Quarterly Journal of Economics* **110** (1995), 445-474.